

**This Page Is Inserted by IFW Operations
and is not a part of the Official Record**

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- **BLACK BORDERS**
- **TEXT CUT OFF AT TOP, BOTTOM OR SIDES**
- **FADED TEXT**
- **ILLEGIBLE TEXT**
- **SKEWED/SLANTED IMAGES**
- **COLORED PHOTOS**
- **BLACK OR VERY BLACK AND WHITE DARK PHOTOS**
- **GRAY SCALE DOCUMENTS**

IMAGES ARE BEST AVAILABLE COPY.

**As rescanning documents *will not* correct images,
please do not report the images to the
Image Problems Mailbox.**

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2001-014326

(43)Date of publication of application : 19.01.2001

(51)Int.Cl.

G06F 17/30

G06F 17/27

G06F 17/21

(21)Application number : 11-183349

(71)Applicant : HITACHI LTD

(22)Date of filing : 29.06.1999

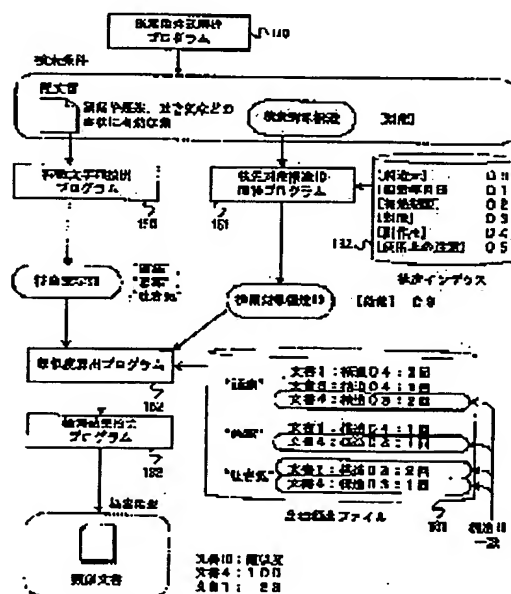
(72)Inventor : MATSUBAYASHI TADATAKA
TADA KATSUMI
SUGAYA NATSUKO
INABA YASUHIKO
YAMAGUCHI AKIHIKO
GOCHI YOSUKE

(54) DEVICE AND METHOD FOR RETRIEVING SIMILAR DOCUMENT BY STRUCTURE SPECIFICATION

(57)Abstract:

PROBLEM TO BE SOLVED: To add the specification of an object structure to be retrieved to retrieval conditions and to improve the retrieval precision when a document which is similar to a seed document (document specified as a retrieval condition) is retrieved.

SOLUTION: A retrieval condition expression analyzing program 130 receives the specification of a seed document and the input of an object structure to be retrieved as retrieval conditions. A featured character string extracting program 150 extracts a featured character string from the text of the specified seed document. A retrieval object structure ID acquiring program 151 converts the specified structure into its ID. A similarity calculating program 152 performs retrieval from an appearance frequency file 181 to acquire the appearance frequency of a document whose structure ID matches the featured character string and calculates the similarity of the similar document based upon the seed document. A retrieval result output program 132 displays the identifier and similarity of the similar document as the retrieval result.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision
of rejection]

[Date of requesting appeal against examiner's
decision of rejection]

[Date of extinction of right]

Copyright (C); 1998,2000 Japan Patent Office

(19)日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11)特許出願公開番号
特開2001-14326
(P2001-14326A)

(43)公開日 平成13年1月19日(2001.1.19)

(51)Int.Cl.⁷

識別記号

F I

テーマコード(参考)

G 0 6 F 17/30
17/27
17/21

G 0 6 F 15/403
15/20
15/40

3 5 0 C 5 B 0 0 9
5 5 0 E 5 B 0 7 5
5 9 0 E
3 4 0
3 7 0 A

審査請求 未請求 請求項の数22 O L (全 16 頁)

(21)出願番号

特願平11-183349

(22)出願日

平成11年6月29日(1999.6.29)

(71)出願人 000005108

株式会社日立製作所

東京都千代田区神田駿河台四丁目6番地

(72)発明者 松林 忠孝

神奈川県川崎市幸区鹿島田890番地 株式

会社日立製作所システム開発本部内

(72)発明者 多田 勝己

神奈川県川崎市幸区鹿島田890番地 株式

会社日立製作所システム開発本部内

(74)代理人 100061893

弁理士 高橋 明夫 (外1名)

最終頁に続く

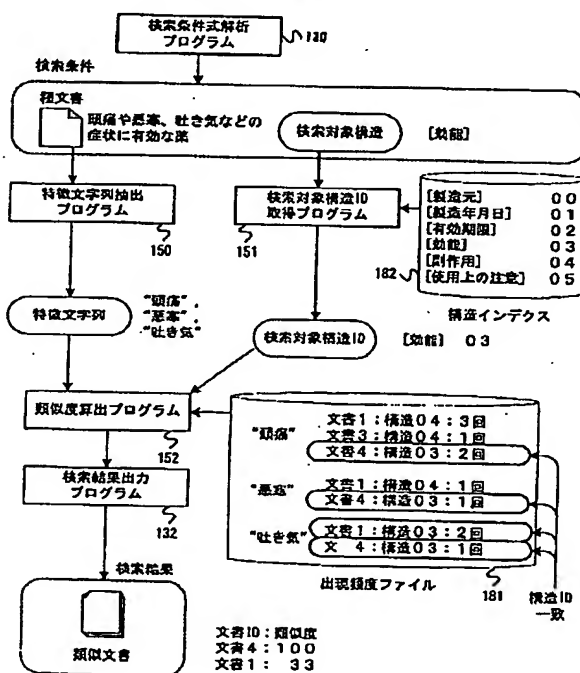
(54)【発明の名称】 構造指定による類似文書の検索装置及び検索方法

(57)【要約】

【課題】 種文書に類似する文書を検索するに際して、検索条件に検索対象構造の指定を付加し、もって検索精度の向上を図る。

【解決手段】 検索条件式解析プログラム130は、検索条件として種文書の指定と検索対象構造の入力を受け、特徴文字列抽出プログラム150は、指定された種文書のテキストから特徴文字列を抽出する。検索対象構造ID取得プログラム151は、指定された構造をそのIDに変換する。類似度算出プログラム152は、出現頻度ファイル181を検索して特徴文字列と構造IDが一致する文書の出現頻度を取得し、類似文書の種文書を基とする類似度を計算する。検索結果出力プログラム132は、検索結果の類似文書の識別子とその類似度を表示する。

図2



【特許請求の範囲】

【請求項 1】 計算機を利用して検索条件として指定された文書あるいは文章（以下まとめて種文書と呼ぶ）に類似する構造化文書を検索する方法であって、類似度計算の検索条件として種文書と該構造化文書に属する少なくとも 1 つの構造の指定を受けるステップと、類似度計算の後、類似度のより高い対象文書を優先して表示するステップとを有することを特徴とする構造指定による類似文書の検索方法。

【請求項 2】 計算機を利用して種文書に類似する構造化文書を検索する方法であって、種文書と検索対象とする構造が指定されたとき、指定された該種文書のテキストから特徴となる文字列を抽出するステップと、抽出された特徴文字列と指定された検索対象構造とが合致する文書を対象として、該特徴文字列に基づく該種文書との類似度を算出するステップと、算出された類似度の高い順に従って表示の優先度を決定するステップとを有することを特徴とする構造指定による類似文書の検索方法。

【請求項 3】 計算機を利用して構造化された種文書に類似する文書を検索する方法であって、種文書と該種文書に属する少なくとも 1 つの構造の指定を受けるステップと、類似度計算の後、類似度のより高い対象文書を優先して表示するステップとを有することを特徴とする構造指定による類似文書の検索方法。

【請求項 4】 計算機を利用して構造化された種文書に類似する文書を検索する方法であって、種文書と検索対象とする構造が指定されたとき、指定された該種文書のテキストのうち指定された構造に属するテキストから特徴となる文字列を抽出するステップと、抽出された特徴文字列と指定された検索対象構造とが合致する文書を対象として該特徴文字列に基づく該種文書との類似度を算出するステップと、算出された類似度の高い順に従って表示の優先度を決定するステップとを有することを特徴とする構造指定による類似文書の検索方法。

【請求項 5】 前記種文書は、表示画面上の指示されたテキストであることを特徴とする請求項 1 又は請求項 2 記載の構造指定による類似文書の検索方法。

【請求項 6】 前記検索対象とする構造が指定される代わりに、前記検索対象から除外されるべき構造が指定されることを特徴とする請求項 1 又は請求項 2 記載の構造指定による類似文書の検索方法。

【請求項 7】 前記種文書に属する構造が指定される代わりに、前記種文書に属し検索対象から除外されるべき構造が指定されることを特徴とする請求項 3 又は請求項 4 記載の構造指定による類似文書の検索方法。

【請求項 8】 検索対象として複数の構造が指定されたとき、前記検索対象構造ごとに類似度を算出し、すべての検索対象構造に亘る類似度の累積値を前記優先度決定の際の最終の類似度とすることを特徴とする請求項 2 又は請求項 4 記載の構造指定による類似文書の検索方法。

【請求項 9】 検索対象として複数の構造が指定されたとき、前記検索対象構造ごとに類似度を算出し、対象とする文書について最も高い類似度を前記優先度決定の際の最終の類似度とすることを特徴とする請求項 2 又は請求項 4 記載の構造指定による類似文書の検索方法。

【請求項 10】 前記特徴文字列を抽出した後に、さらに前記特徴文字列の重要度に応じて採用する特徴文字列を決定することを特徴とする請求項 2 又は請求項 4 記載の構造指定による類似文書の検索方法。

【請求項 11】 前記特徴文字列を抽出した後に、さらに前記指定された種文書に属する構造の重要度に応じて採用する特徴文字列を決定することを特徴とする請求項 4 記載の構造指定による類似文書の検索方法。

【請求項 12】 前記特徴文字列を抽出した後に、さらに前記検索対象構造の重要度に応じて採用する特徴文字列を決定することを特徴とする請求項 2 又は請求項 4 記載の構造指定による類似文書の検索方法。

【請求項 13】 検索結果を表示するに際して、表示する文書について前記検索対象構造ごとに抽出された前記特徴文字列を強調表示することを特徴とする請求項 2 又は請求項 4 記載の構造指定による類似文書の検索方法。

【請求項 14】 検索結果を表示するに際して、表示する文書について前記検索対象構造ごとに前記類似度を表示することを特徴とする請求項 2 又は請求項 4 記載の構造指定による類似文書の検索方法。

【請求項 15】 検索条件として指定された文書あるいは文章（種文書）に類似する構造化文書を検索する装置であって、類似度計算の検索条件として種文書と該構造化文書に属する少なくとも 1 つの構造の指定を受ける手段と、類似度計算の後、類似度のより高い対象文書を優先して表示する手段とを有することを特徴とする構造指定による類似文書の検索装置。

【請求項 16】 種文書に類似する構造化文書を検索する装置であって、種文書と検索対象とする構造が指定されたとき、指定された該種文書のテキストから特徴となる文字列を抽出する手段と、抽出された特徴文字列と指定された検索対象構造とが合致する文書を対象として、該特徴文字列に基づく該種文書との類似度を算出する手段と、算出された類似度の高い順に従って表示の優先度を決定する手段とを有することを特徴とする構造指定による類似文書の検索装置。

【請求項 17】 構造化された種文書に類似する文書を検索する装置であって、種文書と該種文書に属する少なくとも 1 つの構造の指定を受ける手段と、類似度計算の後、類似度のより高い対象文書を優先して表示する手段とを有することを特徴とする構造指定による類似文書の検索装置。

【請求項 18】 構造化された種文書に類似する文書を検索する装置であって、種文書と検索対象とする構造が指定されたとき、指定された該種文書のテキストのうち指

定された構造に属するテキストから特徴となる文字列を抽出する手段と、抽出された特徴文字列と指定された検索対象構造とが合致する文書を対象として該特徴文字列に基づく該種文書との類似度を算出する手段と、算出された類似度の高い順に従って表示の優先度を決定する手段とを有することを特徴とする構造指定による類似文書の検索装置。

【請求項19】 計算機読み取り可能な記憶媒体に格納されたプログラムであって、該プログラムは、検索条件として指定された文書あるいは文章（種文書）に類似する構造化文書を検索するプログラムであり、類似度計算の検索条件として種文書と該構造化文書に属する少なくとも1つの構造の指定を受けるプログラム手段と、類似度計算の後、類似度のより高い対象文書を優先して表示するプログラム手段とを有することを特徴とするプログラムを格納する記憶媒体。

【請求項20】 計算機読み取り可能な記憶媒体に格納されたプログラムであって、該プログラムは、種文書に類似する構造化文書を検索するプログラムであり、種文書と検索対象とする構造が指定されたとき、指定された該種文書のテキストから特徴となる文字列を抽出するプログラム手段と、抽出された特徴文字列と指定された検索対象構造とが合致する文書を対象として、該特徴文字列に基づく該種文書との類似度を算出するプログラム手段と、算出された類似度の高い順に従って表示の優先度を決定するプログラム手段とを有することを特徴とするプログラムを格納する記憶媒体。

【請求項21】 計算機読み取り可能な記憶媒体に格納されたプログラムであって、該プログラムは、構造化された種文書に類似する文書を検索するプログラムであり、種文書と該種文書に属する少なくとも1つの構造の指定を受けるプログラム手段と、類似度計算の後、類似度のより高い対象文書を優先して表示するプログラム手段とを有することを特徴とするプログラムを格納する記憶媒体。

【請求項22】 計算機読み取り可能な記憶媒体に格納されたプログラムであって、該プログラムは、構造化された種文書に類似する文書を検索するプログラムであり、種文書と検索対象とする構造が指定されたとき、指定された該種文書のテキストのうち指定された構造に属するテキストから特徴となる文字列を抽出するプログラム手段と、抽出された特徴文字列と指定された検索対象構造とが合致する文書を対象として該特徴文字列に基づく該種文書との類似度を算出するプログラム手段と、算出された類似度の高い順に従って表示の優先度を決定するプログラム手段とを有することを特徴とするプログラムを格納する記憶媒体。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】 本発明は、検索条件として指

定された文書（種文書）に類似する文書を検索する装置及び方法に係わり、特に構造化文書の構造を対象として検索を行う装置及び方法に関する。

【0002】

【従来の技術】 近年、パーソナルコンピュータやインターネット等の普及に伴い、データベースに蓄積される電子化文書の数が増大しており、膨大な電子化文書の中からユーザが所望する情報を含んだ文書を検索精度よく、高速かつ効率的に検索したいという要求が高まっている。

【0003】 このような要求に対して種々の検索技術が提案されている。例えば特開平10-240752号公報によれば、文書を構成する個々の論理的な構造要素が識別できる文書（以下、構造化文書と呼ぶ）を対象として、論理構造に関する条件を検索条件中に付加した検索を行うことにより、精度の高い検索を行うことができる。

【0004】 また特開平11-143902号公報は、ユーザが自分の所望する内容の文書あるいは文章（以下、種文書と呼ぶ）を指定し、その文書と類似する文書を検索する類似文書検索技術を開示する。この技術によれば、サンプル文書を示すだけで目的の文書を簡単に検索でき、ユーザが複雑な検索条件式を考えたり入力する手間が省け、効率的な検索ができる。

【0005】

【発明が解決しようとする課題】 上記の特開平11-143902号公報の技術によれば、ユーザは使い勝手がよく効率的な検索ができるが、以下に例示するように検索精度の問題を残している。

【0006】 図3は、従来の類似文書検索システムの処理手順を示す図である。検索条件取得プログラムは、検索条件を入力するためのガイダンス画面を表示装置上に表示する。例えば種文書を含む複数の候補文書の文書番号や見出しなどの一覧情報を表示する。検索条件として種文書が指定されると、特徴n-gram抽出プログラムが起動され、文書ファイルから種文書のテキスト全文を取り出し、テキスト中から特徴文字列を抽出する。次に類似度算出プログラムが起動され、特徴文字列に対応して文書番号とその特徴文字列の出現回数が登録してある出現頻度ファイルを参照し、種文書の特徴文字列に基づいて同じ特徴文字列を使用する関連文書の種文書に対する類似度を算出して候補文書の文書番号、類似度、見出しなどの一覧情報を検索結果として表示する。

【0007】 図3の検索結果によれば、種文書とよく類似する文書は文書4であるにもかかわらず、特徴文字列の出現頻度がより高いために関連の薄い文書1の類似度がより高くなり、優先的に表示されるという問題がある。

【0008】 本発明の目的は、複雑な検索条件の入力を避けるが検索精度のよい類似文書の検索装置及び検索方

法を提供することにある。

【0009】

【課題を解決するための手段】本発明は、計算機を利用して種文書に類似する構造化文書を検索する方法であって、類似度計算の検索条件として種文書と構造化文書に属する少なくとも1つの構造の指定を受けるステップと、類似度計算の後、類似度のより高い対象文書を優先して表示するステップとを有する構造指定による類似文書の検索方法を特徴とする。

【0010】また本発明は、構造化された種文書に類似する文書を検索する方法であって、種文書とその種文書に属する少なくとも1つの構造の指定を受けるステップと、類似度計算の後、類似度のより高い対象文書を優先して表示するステップとを有する構造指定による類似文書の検索方法を特徴とする。

【0011】さらに本発明は、上記の機能を備える検索装置を特徴とする。

【0012】なおここで構造を指定するとは、文書を構成する論理的な構造要素の名称を指定することを意味する。

【0013】

【発明の実施の形態】以下、本発明の実施形態について図面を用いて説明する。

【0014】図1は、第1の実施形態の類似文書検索システムの構成図である。本システムを実現する計算機ハードウェアは、表示装置100、入力装置101、中央処理装置(CPU)102、外部記憶装置103、フロッピーディスクドライブ(FDD)104、主メモリ106とこれら装置間を接続するバス107から構成される。

【0015】外部記憶装置103は、テキスト180、出現頻度ファイル181及び構造インデクス182を格納する。テキスト180は、構造化文書ファイルあるいは構造化されていない文書ファイルの集合を格納する。ここで構造化文書とは、SGML、XMLなどの標準形式に準拠した論理構造をもつ文書、あるいは各論理構造ごとに抽出された複数のフラットテキストから構成されるものである。FDD104を介してフロッピーディスク105に格納されている文書が主メモリ106を経由してテキスト180に登録される。

【0016】主メモリ106に格納されるシステム制御プログラム110は、オペレーティングシステム、グラフィカル・ユーザインタフェースを提供するプログラムなどを含む。文書登録制御プログラム111は、文書登録用のプログラムの実行を制御する。登録プログラムには、テキスト登録プログラム120、出現頻度計数プログラム140を含む出現頻度ファイル作成プログラム121及び構造インデクス作成プログラム122がある。テキスト登録プログラム120は、フロッピーディスク105上の文書をテキスト180に登録するプログラム

である。

【0017】検索制御プログラム112は、類似文書の検索に係わるプログラムの実行を制御するプログラムである。検索用のプログラムには、検索条件式解析プログラム130、類似文書検索プログラム131及び検索結果出力プログラム132がある。類似文書検索プログラム131には、特徴文字列抽出プログラム150、検索対象構造ID取得プログラム151及び類似度算出プログラム152が含まれる。これら検索用プログラムの機能については、以下の検索処理手順の説明の中で説明する。検索制御プログラム112及び検索用プログラムを記憶媒体に格納し、駆動装置を介して主メモリ106に読み込み、CPU102によって実行することが可能である。

【0018】主メモリ106中に格納される共有ライブラリ160として、構造化文書解析プログラム170がある。またワークエリア161は、テキスト180、出現頻度ファイル181、構造インデクス182から読み込んだデータ等の一時記憶領域や作業用領域として使用される領域である。

【0019】出現頻度ファイル181は、図2の一部に示すように文字列又は単語に対応して、文書番号、その文字列が含まれる論理構造のID及びその論理構造中の出現回数を格納する。出現頻度ファイル181に登録される対象文書は、構造化文書または構造化していない文書である。出現頻度ファイル作成プログラム121は、テキスト180中の文書を1つずつ読み込み、テキスト中から特徴文字列を抽出し、出現頻度計数プログラム140によって各論理構造ごとの特徴文字列の出現回数を計数し、出現頻度ファイル181を作成して外部記憶装置103に登録する。構造化していない文書については、論理構造の区分が指定されると、その指定に従って特徴文字列と論理構造IDとを対応づける。例えば特開平11-143902号公報は出現頻度ファイル作成プログラムの処理手順を開示する。

【0020】構造インデクス182は、図2の一部に示すように論理構造とそのIDの対応関係を格納する。構造インデクス作成プログラム122は、構造化文書解析プログラム170を呼び出し、テキスト180から読み込んだ文書テキストの論理構造を解析して各論理構造にIDを付与して外部記憶装置103に登録する。例えば特開平10-240752号公報は、構造インデクス作成プログラムの処理手順を開示する。

【0021】図2は、第1の実施形態の処理手順を示す図である。第1の実施形態では、種文書が構造化されていない文書、検索対象文書が出現頻度ファイル181に登録済みの文書(構造化文書又は構造化されていない文書)とする。検索条件式解析プログラム130は、表示装置100上にガイダンス画面を表示し、検索条件式の入力を受け付ける。

【0022】ここで検索条件式は、種文書及び少なくとも1つの検索対象構造である。種文書はすでに見出し等が表示された複数の文書候補のうちの1つを選択することが可能であるし、入力装置101から直接入力することも可能であるし、FDD104やCD-ROM装置（図には示していない）、ネットワーク（図には示していない）等を介して入力することも可能である。

【0023】さらに図12に示すように、表示装置100に種文書入力用領域1200、検索対象構造入力用領域1201および検索実行ボタン1202を備えた画面インタフェースを介して検索条件式が入力されるものとしてもよい。種文書入力用領域1200には、入力装置101より種文書を直接入力することも可能であるし、あるいは検索結果表示画面（図には示していない）上のテキストを種文書入力用領域1200にコピーすることも可能である。あるいは種文書はこのようなテキストのうちの指示された部分であってもよい。

【0024】また検索対象構造は、表示されるドロップダウンメニューから少なくとも1つを選択することが可能である。複数の検索対象構造が指定された場合の検索条件では、各構造に対して重みを付与することが可能である。ここで重みは、重み入力用領域（図には示していない）を介して入力されるものでもよいし、システム定義ファイル（図には示していない）で定義されるものとしてもよい。

【0025】入力装置101を介して種文書及び検索対象構造が入力されると、検索条件式解析プログラム130は指定された検索条件から種文書のテキストを取得する。なお検索対象とする構造を指定する代わりに、検索対象としない構造（検索対象から除外する構造）を指定してもよい。その場合には、検索条件式解析プログラム130は、残りの構造を検索対象構造とする。また検索対象構造を検索条件式の1つとして入力装置101を介して入力する代わりにあらかじめシステム定義ファイル（図には示していない）に設定された検索対象構造を用いてもよい。

【0026】検索対象構造ID取得プログラム151は、構造インデクス182を参照して指定された構造に対応する識別子を検索対象構造IDとして取得する。また特徴文字列抽出プログラム150は、テキスト180から指定された種文書のテキスト全文を取り出し、特徴文字列を抽出し、抽出した特徴文字列の出現回数を計数する。特徴文字列の抽出方法としては、例えば特開平11-143902号公報に記載された方法を用いることができる。図2の例では抽出した特徴文字列のうち優先度の高いものを採用している。あるいは文書テキストから単語を切り出し、単語辞書（図には示していない）を参照して登録された単語との一致をチェックしながら単語を抽出してもよい。

【0027】次に類似度算出プログラム152は、出現

頻度ファイル181を参照して抽出された各特徴文字列又は単語と検索対象構造IDが一致する文書の文書番号とその出現頻度を取得する。次に出現頻度ファイル181を参照して取得した各文書の検索対象構造IDについて抽出された特徴文字列又は単語以外の他の文字列又は単語の出現頻度を取得し、各文書ごとに種文書との類似度を算出する。類似度算出方法としては、例えば特開平11-143902号公報に記載の数式1を用いることができる。あるいは種文書の各特徴文字列（単語）の正規化された出現ウェイトを要素とする特徴ベクトルと取得した各文書の特徴ベクトルを求め、種文書と他文書の特徴ベクトルの内積によって各文書の類似度を計算してもよい。

【0028】最後に検索結果出力プログラム132は、取得した文書を類似度の高い順に並べ替え、類似度の高い順に従って表示の優先度を決定し、優先度の高い文書から順に文書番号とその類似度を表示装置100上に表示する。ファイル（図には示していない）を参照して各文書の見出し、概要などの書誌事項を取得して併せて表示してもよい。類似文書との比較のために種文書の文書番号、類似度、見出しなどを併せて表示することも可能である。

【0029】なお複数の検索対象構造が指定された場合に、各文書の類似度を算出するに際して、各検索対象構造の類似度を全体に亘って累積する累積値を求め、文書をこの類似度の累積値の大きい順に並べ替えてもよい。ここで累積値とは、各検索対象構造ごとの類似度の総和、2乗和の平方根を求めたものなどである。あるいは各文書について複数の検索対象構造の各類似度のうち最も高い類似度を採用し、文書をこの採用した類似度の大きい順に並べ替えてもよい。例えば特許明細書中の「請求項n」のように文書中に同一種類の論理構造が繰り返し出現する場合に、各論理構造ごとに類似度を算出し、その中で最も高い類似度を採用して種文書の類似度と比較すると、同一種類の論理構造の順番には無関係に内容の類似度の高い論理構造同志の比較をすることができる。また類似度の累積値を求めるモードと、最も高い類似度を採用するモードの両方を設け、検索条件の1つとしていずれかのモードを選択できるようにしてもよいし、あらかじめシステム定義ファイル（図には示していない）に選択するモードを設定できるようにしてもよい。

【0030】図4は、第2の実施形態の類似文書検索プログラム131aの構成を示す図である。第2の実施形態では、特徴文字列抽出プログラム150aに種文書構造解析プログラム400が加わっている。種文書構造解析プログラム400は、共有ライブラリ160に格納されている構造化文書解析プログラム170を呼び出す構成をとる。また類似度算出プログラム152aに対応構造判定プログラム401が加わっている。

【0031】図5は、第2の実施形態の処理手順を示す図である。第2の実施形態では種文書が構造化文書、検索対象文書が出現頻度ファイル181に登録済みの文書（構造化文書又は構造化されていない文書）とする。第2の実施形態の検索条件は種文書及び種文書に属する少なくとも1つの構造である。入力装置101を介して種文書及び構造が入力されると、検索条件式解析プログラム130は指定された検索条件から種文書のテキストを取得する。検索条件で指定された種文書の論理構造と検索対象文書の論理構造が一致するものとする。なお検索対象とする構造を指定する代わりに、検索対象としない構造（検索対象から除外する構造）を指定してもよい。その場合には、検索条件式解析プログラム130は、種文書に属する残りの構造を検索対象構造とする。第1の実施形態と同様に検索対象構造をあらかじめシステム定義ファイルに設定しておいてもよい。

【0032】次に種文書構造解析プログラム400は、テキスト180から種文書のテキストを取り出し、種文書の構造を解析して指定された構造に関する本文テキストのみを抽出する。種文書のテキストに指定された構造が含まれていないときにはエラーとする。文書の構造解析の方法としては、例えば特開平10-240752号公報に文書構造解析プログラムの処理手順として記載されている。次に検索対象構造ID取得プログラム151は、構造インデックス182を参照して指定された構造に対応する検索対象構造IDを取得する。また特徴文字列抽出プログラム150aは、抽出されたテキストの特徴文字列（単語）を抽出し、抽出した特徴文字列の出現回数を計数する。なお種文書について、すでに各構造の特徴文字列が抽出され、その構造ごとの出現回数が計数されており、出現頻度ファイル181のように登録されているのであれば、そのファイルを参照して指定された構造、特徴文字列と出現回数を抽出するだけでよい。この場合には種文書構造解析プログラム400及び特徴文字列抽出プログラム150aの処理をスキップできる。

【0033】次に類似度算出プログラム152aは、第1の実施形態と同様に出現頻度ファイル181を参照して抽出された各特徴文字列（単語）と検索対象構造IDが一致する文書の文書番号を取得し、各文書ごとに種文書との類似度を算出する。この際に対応構造判定プログラム401は、種文書の構造と、出現頻度ファイル181から取得された文書の構造IDとの対応をとり、特徴文字列を検索対象構造ごとのグループに分け、検索対象構造ごとの類似度を算出する。複数の検索対象構造が指定された場合に、各文書の最終的な類似度を算出する方法は第1の実施形態と同様である。最後に検索結果出力プログラム132は、取得した文書を算出した類似度の高い順に並べ替えてその文書番号、類似度、見出し等を表示装置100上に表示する。

【0034】なお上記の第2の実施形態の説明では、指

定された種文書の構造と検索対象構造とが一致するものとしたが、両者が別の論理構造であってもよい。すなわち種文書の構造が指定され、これとは別の検索対象構造が指定された場合、種文書構造解析プログラム400及び特徴文字列抽出プログラム150aは、指定された種文書の構造に注目して特徴文字列を抽出し、類似度算出プログラム152aは指定された検索対象構造のIDに注目して検索対象文書を検索する。また対応構造判定プログラム401は、指定された種文書の構造と指定された検索対象構造が同一グループとみなして対応づけをする。例えば薬の効能書の「副作用」を特徴文字列を抽出するときの対象構造とし、「効能」を検索対象文書の類似度を計算するときの検索対象構造とすることにより、種文書に記載の薬のもつ副作用を抑える薬について記載された文書を探し出すことが可能となる。

【0035】図6は、第2の実施形態の問題点を説明する図である。この例では種文書に関する構造として「効能」「副作用」「使用上の注意」が指定され、これらの構造が検索対象構造とみなされ、検索が実行されている。その結果「服用」「自動車」「運転」など薬の効能書にとって重要度が小さいか無意味な特徴文字列が抽出され、これらの特徴文字列を含む特徴文字列に基づく類似度算出の結果として、文書2、文書3などあまり重要でない文書の類似度が無視できない程の値を示し、検索結果として文書2、文書3などが挙がったことを示している。

【0036】図7は、第3の実施形態の類似文書検索プログラム131bの構成を示す図である。第3の実施形態では、特徴文字列抽出プログラム150bにさらに構造重みプログラム600が加わっている。

【0037】図8は、第3の実施形態の処理手順を示す図である。第3の実施形態は、特徴文字列抽出プログラム150aが抽出した特徴文字列に対して構造重みプログラム600を適用する以外は第2の実施形態の処理と同じである。構造重みプログラム600は、論理構造ごとに重要度が設定してあるシステム定義ファイルを参照して、各論理構造ごとにその重要度に応じて抽出した特徴文字列の中から検索用として採用する特徴文字列の数を決定する。例えば「重要」の構造は抽出されたすべての特徴文字列を採用し、「普通」の構造は抽出された特徴文字列の重要度に従って一部の特徴文字列のみを採用する。あるいは各論理構造ごとに採用する特徴文字列の数をシステム定義ファイルに設定し、抽出された特徴文字列からその重要度が上位の所定数の特徴文字列を採用してもよい。また所定の文字種からそれぞれ所定数を採用するようにしてもよい。特徴文字列の重要度を算出する方法としては、例えば特開平11-143902号公報は数式2として特徴文字列の重要度の算出式を挙げている。なお各論理構造の重要度や特徴文字列の採用個数をシステム定義ファイルに設定する代わりに、検索条件

式の一部として入力装置101を介して指定してもよい。なお論理構造の重要度あるいは特徴文字列の重要度により採用する特徴文字列を決定する方式は、上記の第1の実施形態にも適用可能である。

【0038】以上のようにして特徴文字列を絞り込んだ上で類似度算出プログラム152a及び対応構造判定プログラム401を適用すると、検索結果から重要度の少ない文書を排除することができる。また第1、第2の実施形態に比べて特徴文字列の数が削減されることになるので、出現頻度ファイル181を検索する際の検索時間を短縮できる。

【0039】図9は、第4の実施形態の検索結果表示プログラム132aの構成を示す図である。第4の実施形態では、検索結果表示プログラム132aに構造別表示方法取得プログラム700が加わっている。

【0040】図10は、第4の実施形態の処理手順を示す図である。第4の実施形態は、検索結果表示プログラム132aの処理を除いては第1～第3の実施形態の処理と同じである。構造別表示方法取得プログラム700は、類似度算出プログラム152aの処理結果として挙げられた文書について検索対象構造別の類似度を表示する。また検索対象構造ごとに抽出された特徴文字列を強調表示する。

【0041】図11は、構造別表示方法取得プログラム700の処理手順を示すPAD図である。構造別表示方法取得プログラム700は、特徴文字列抽出プログラム150aにより抽出された各構造ごとの特徴文字列をそれぞれワークエリア161に格納する（ステップ701）。次に類似度算出プログラム152aにより算出された各構造ごとの類似度をワークエリア161に格納する（ステップ702）。次に検索された各文書の指定されたすべての構造について以下の処理を繰り返す（ステップ703）。まずワークエリア161に格納された当該構造の類似度を取得し表示する（ステップ704）。次にワークエリア161に格納された当該構造の特徴文字列を取得し、強調表示する（ステップ705）。なおこの実施形態では各論理構造ごとに類似度と特徴文字列の強調表示とを行うものとしたが、いずれか一方のみを行ってもよい。また検索結果の表示条件をシステム定義ファイル上に設定してもよいし、検索条件式の一部として指定してもよい。

【0042】なお種文書及び検索結果として挙げられた類似文書について、各々文書番号に対応して見出し、概要などを表示し、これらのテキストに含まれ、採用された特徴文字列を強調表示してもよい。このように表示すると、類似文書に含まれる特徴文字列を種文書に含まれる特徴文字列と比較することができる。

【0043】なお上記実施形態で使用した出現頻度ファイル181の代わりに特開平11-143902号公報のn-gramインデックスを用いてもよい。すなわち特

開平11-143902号公報と特開平10-240752号公報の構造化された文字列インデックスを組み合わせると、出現頻度ファイル181に代わるファイルを構成可能である。同一文字列についての1つ以上の出現位置はその文字列の出現回数をも示している。

【0044】なお上記第1～第4の実施形態では種文書として1つの文書が指定されるものとしたが、複数の種文書を指定できるものとしてもよい。ここで特徴文字列としては、それぞれの種文書から抽出された特徴文字列をすべて用いるものとしてもよいし、それぞれの種文書に共通して含まれる特徴文字列を用いるものとしてもよい。

【0045】

【発明の効果】以上述べたように本発明によれば、類似文書検索の検索条件として論理構造の指定を付加するので、類似文書検索の利点を最大限に生かしながら検索精度を高めることができる。なお複数の検索対象構造が指定される場合に、あらかじめ設定された論理構造の重要度または特徴文字列の重要度に応じて関連の薄い特徴文字列を排除でき、さらに検索精度を高めることができる。また種文書は、構造化文書と構造化していない文書のいずれも可能であり、ユーザが種文書の選択に注意を払う必要がない。

【図面の簡単な説明】

【図1】実施形態の類似文書検索システムの構成図である。

【図2】第1の実施形態の処理手順を示す図である。

【図3】従来の類似文書検索システムの処理手順を示す図である。

【図4】第2の実施形態の類似文書検索プログラム131aの構成を示す図である。

【図5】第2の実施形態の処理手順を示す図である。

【図6】第2の実施形態の問題点を説明する図である。

【図7】第3の実施形態の類似文書検索プログラム131bの構成を示す図である。

【図8】第3の実施形態の処理手順を示す図である。

【図9】第4の実施形態の検索結果出力プログラム132aの構成を示す図である。

【図10】第4の実施形態の処理手順を示す図である。

【図11】第4の実施形態の構造別表示方法取得プログラムの処理手順を示す図である。

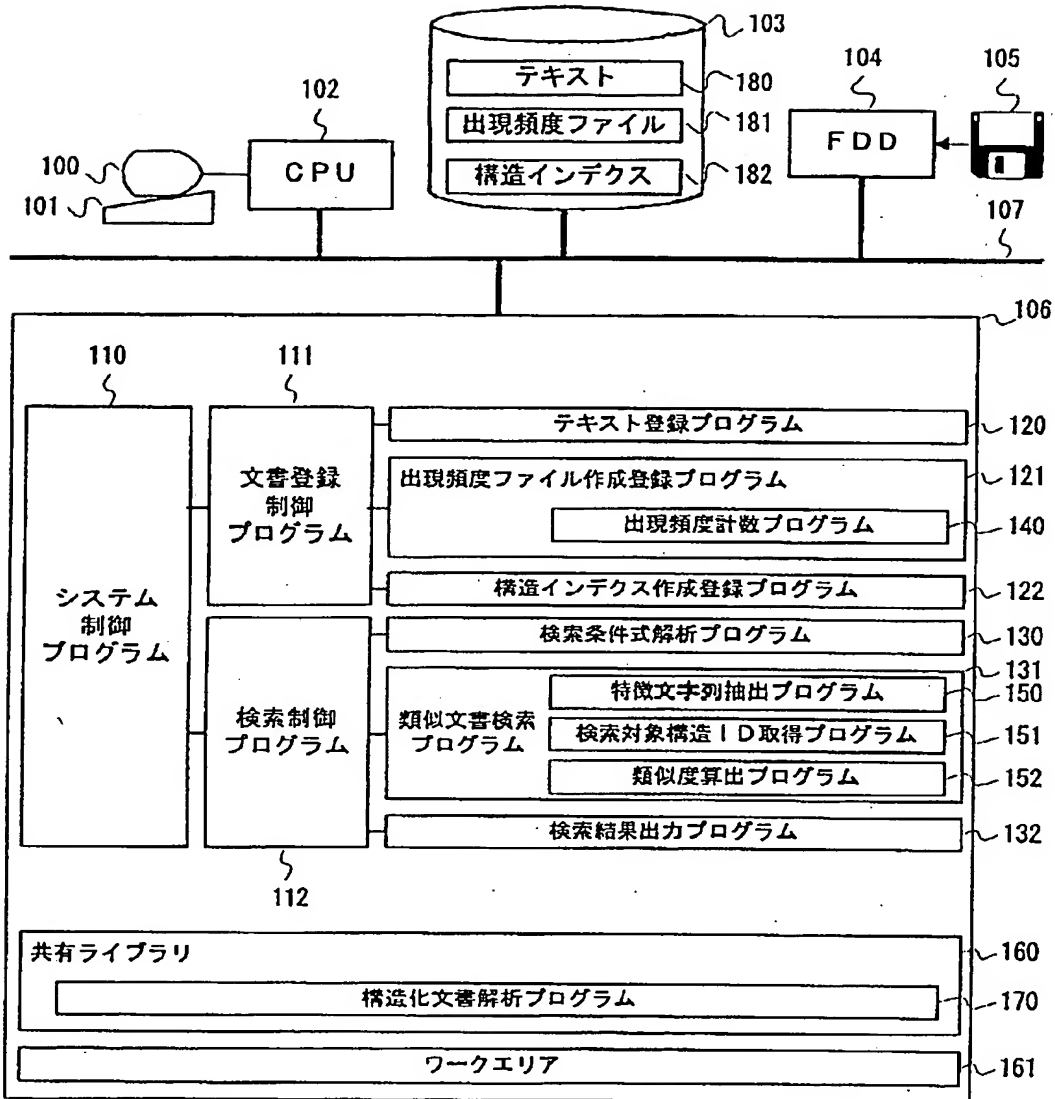
【図12】検索条件入力画面の例を示す図である。

【符号の説明】

131：類似文書検索プログラム、132：検索結果出力プログラム、150：特徴文字列抽出プログラム、151：検索対象構造ID取得プログラム、152：類似度算出プログラム、180：テキスト、181：出現頻度ファイル、182：構造インデックス、400：種文書構造解析プログラム

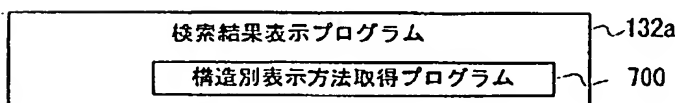
【図1】

図1



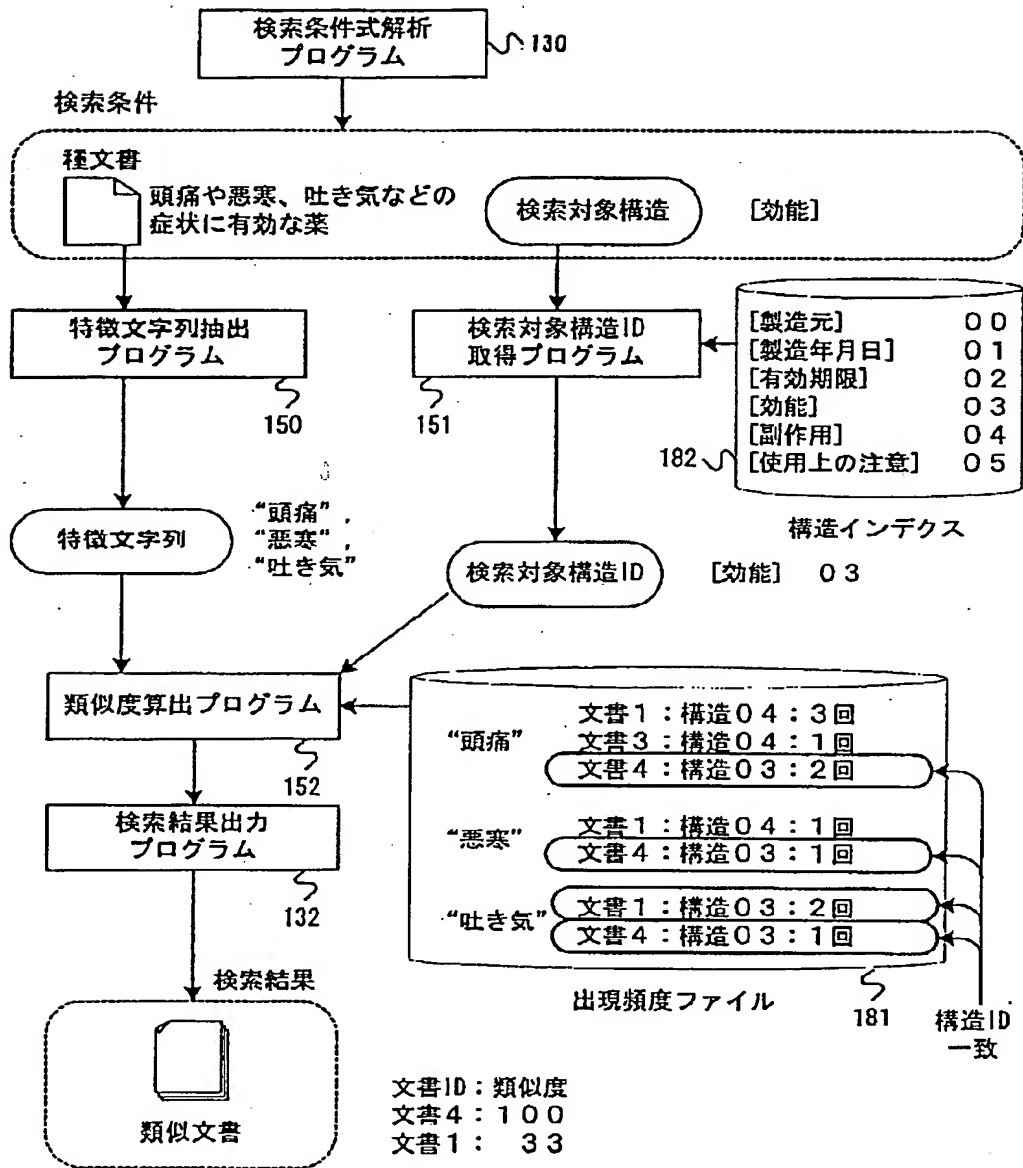
【図9】

図9



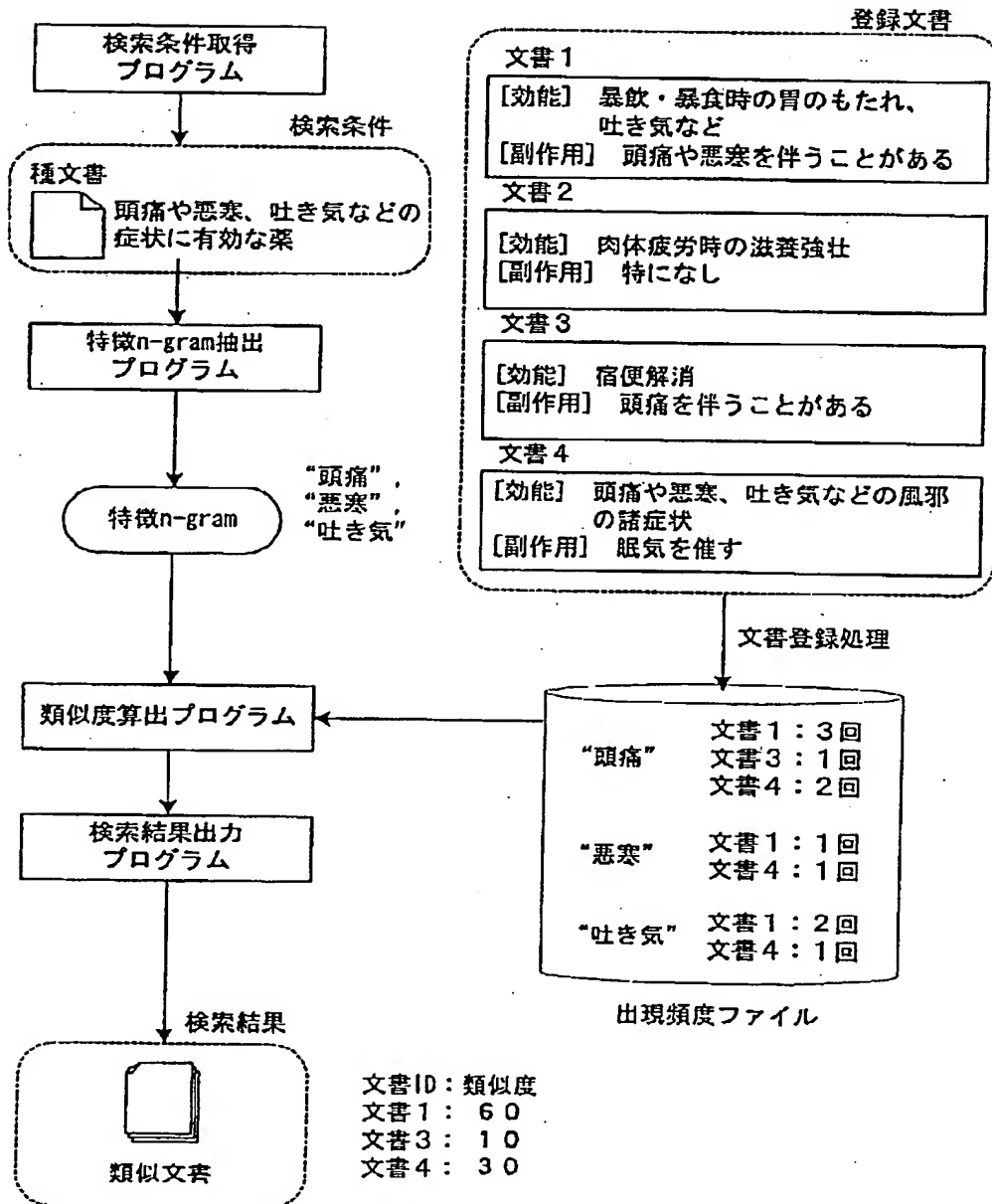
【図2】

図2



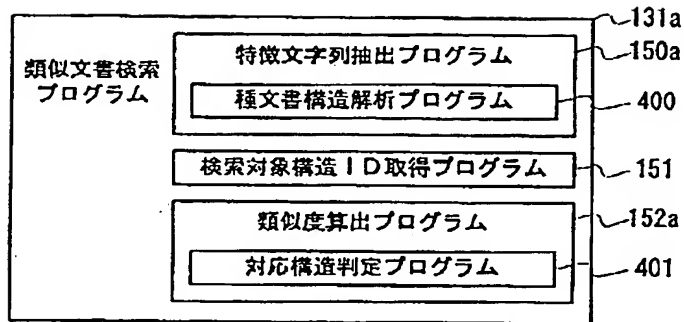
【図3】

図3



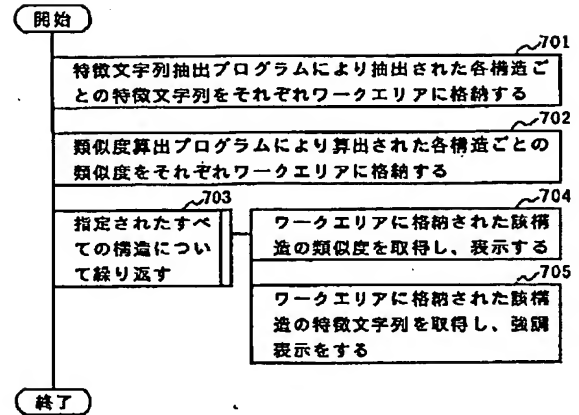
【図 4】

図 4



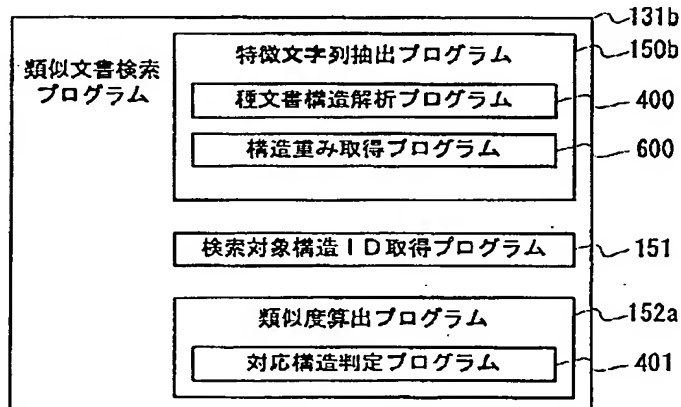
【図 11】

図 11



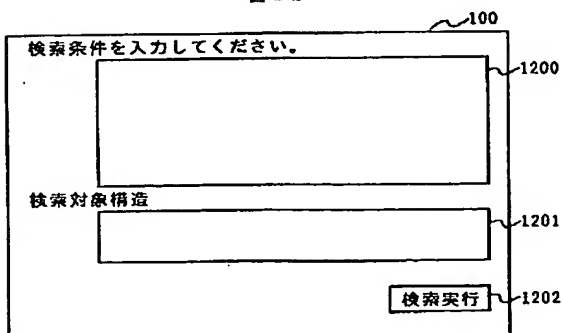
【図 7】

図 7



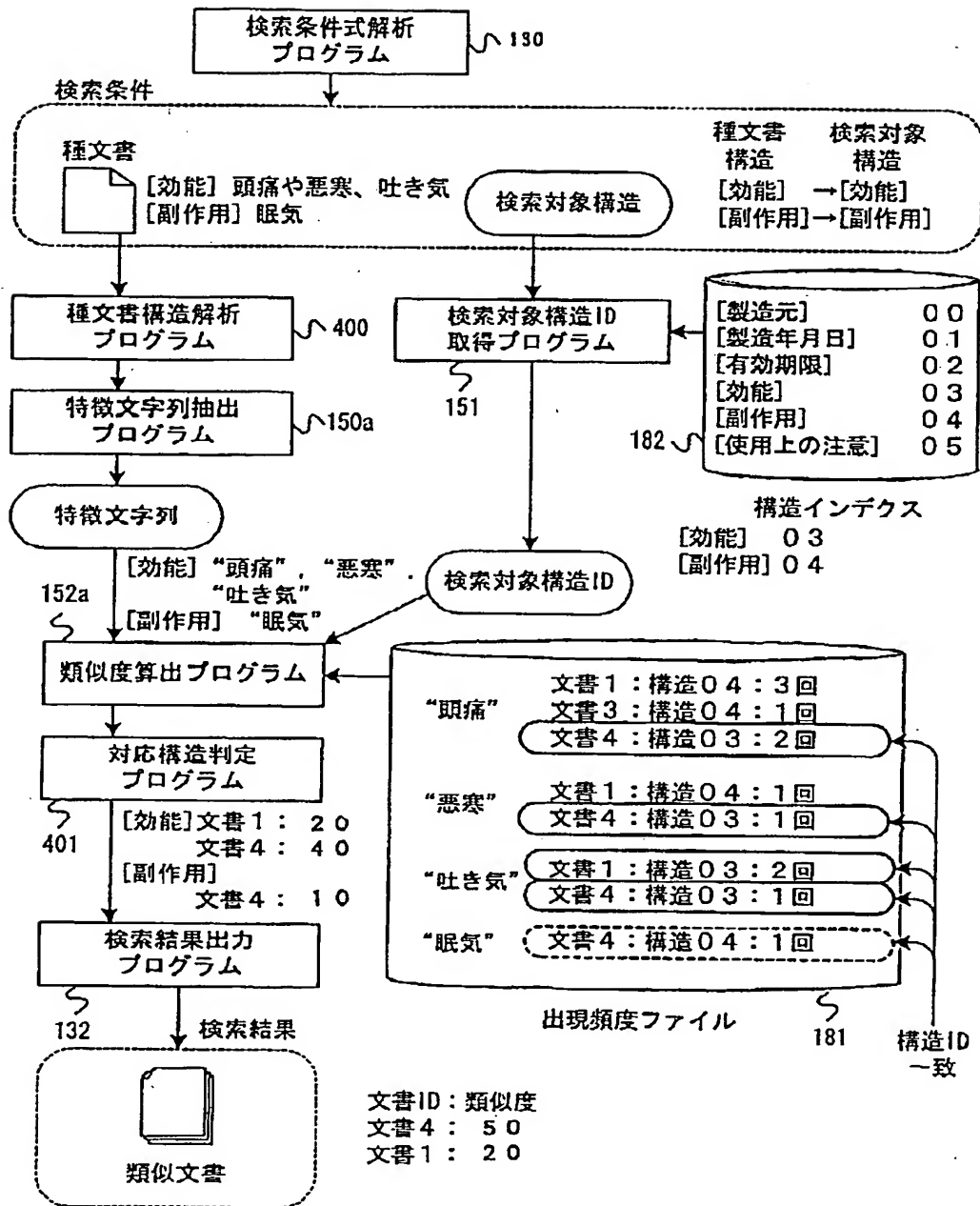
【図 12】

図 12

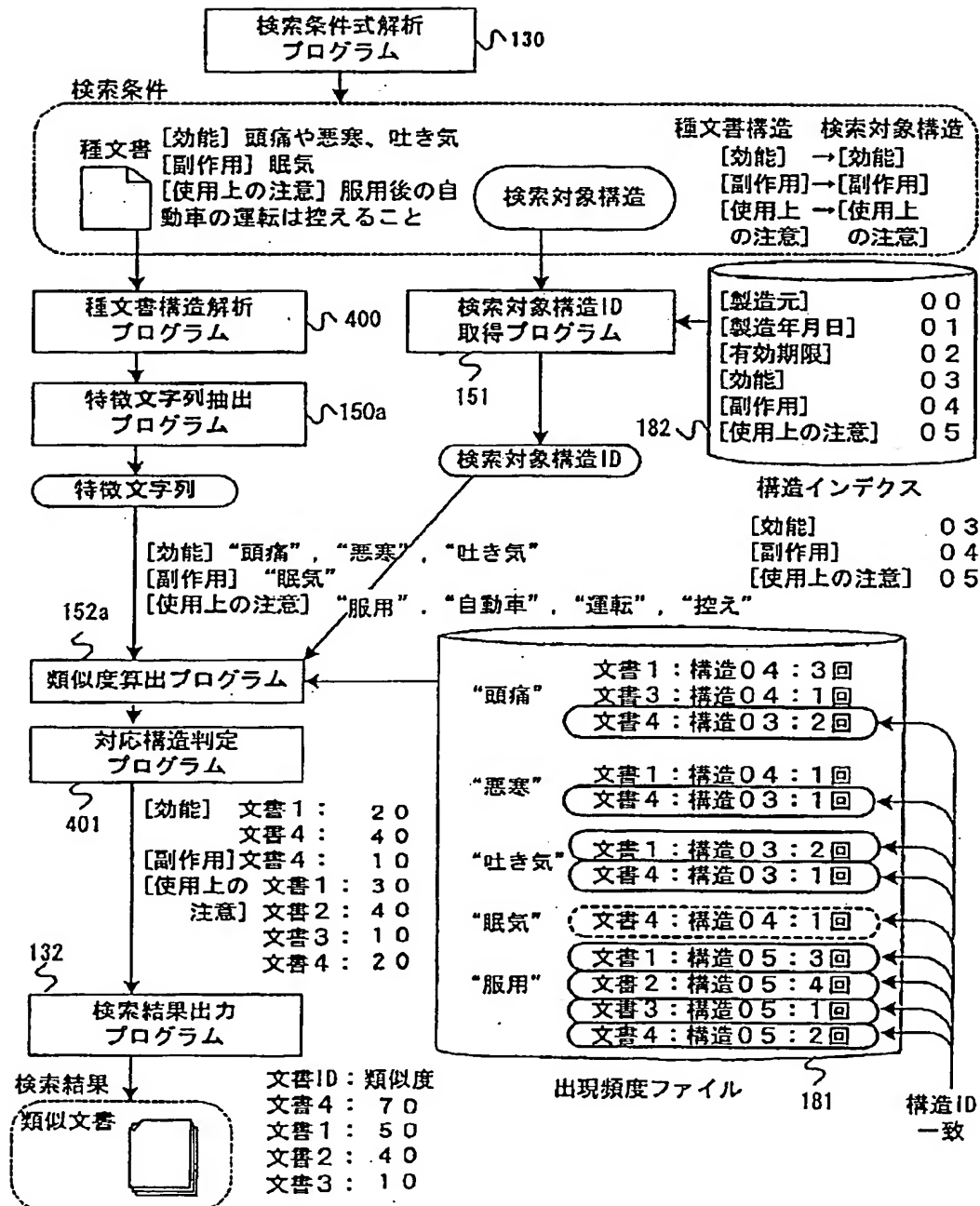


【図5】

図5

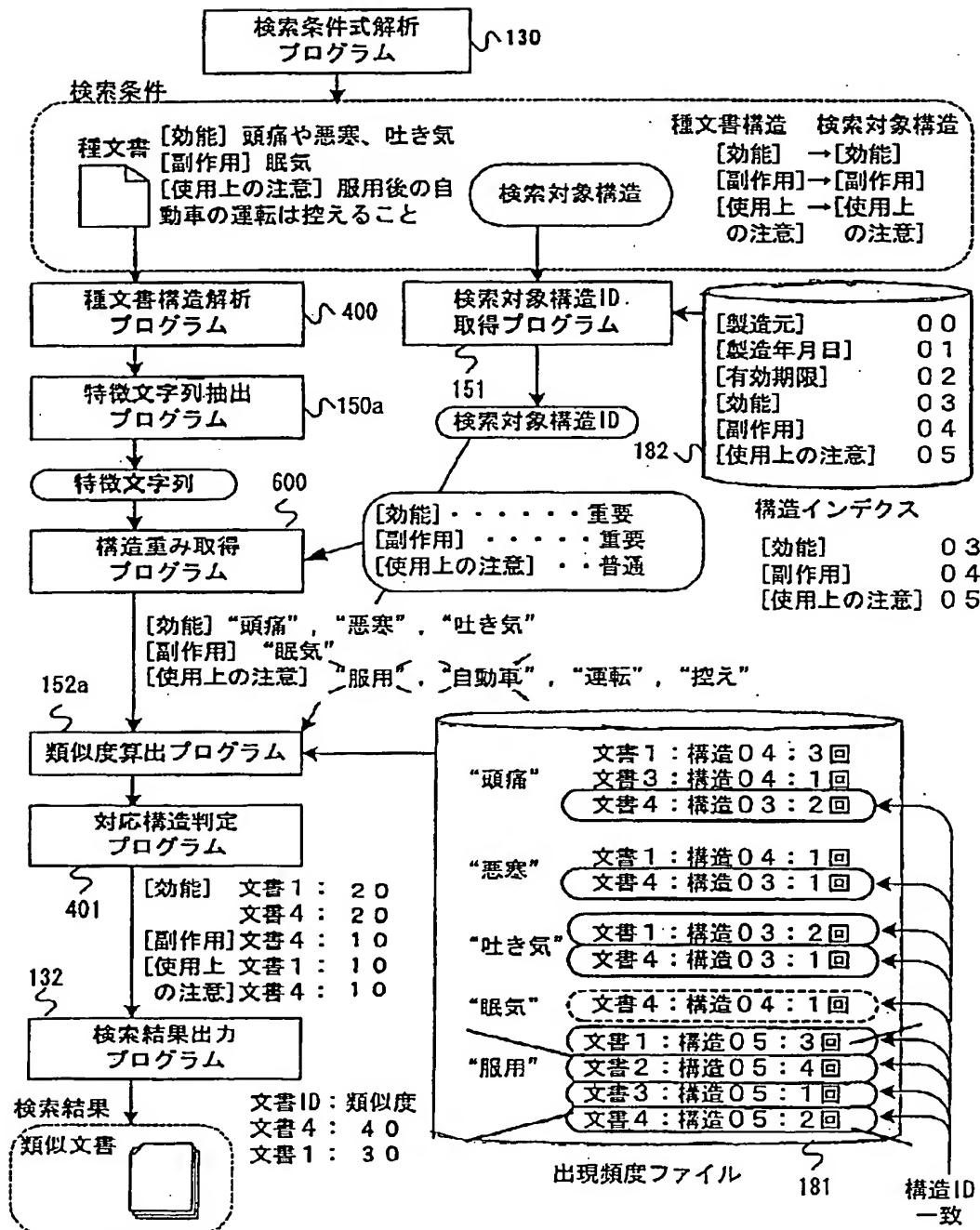


Number of people	Time taken (minutes)
1	100
2	80
3	65
4	55
5	48
6	42
7	38
8	35
9	32
10	30



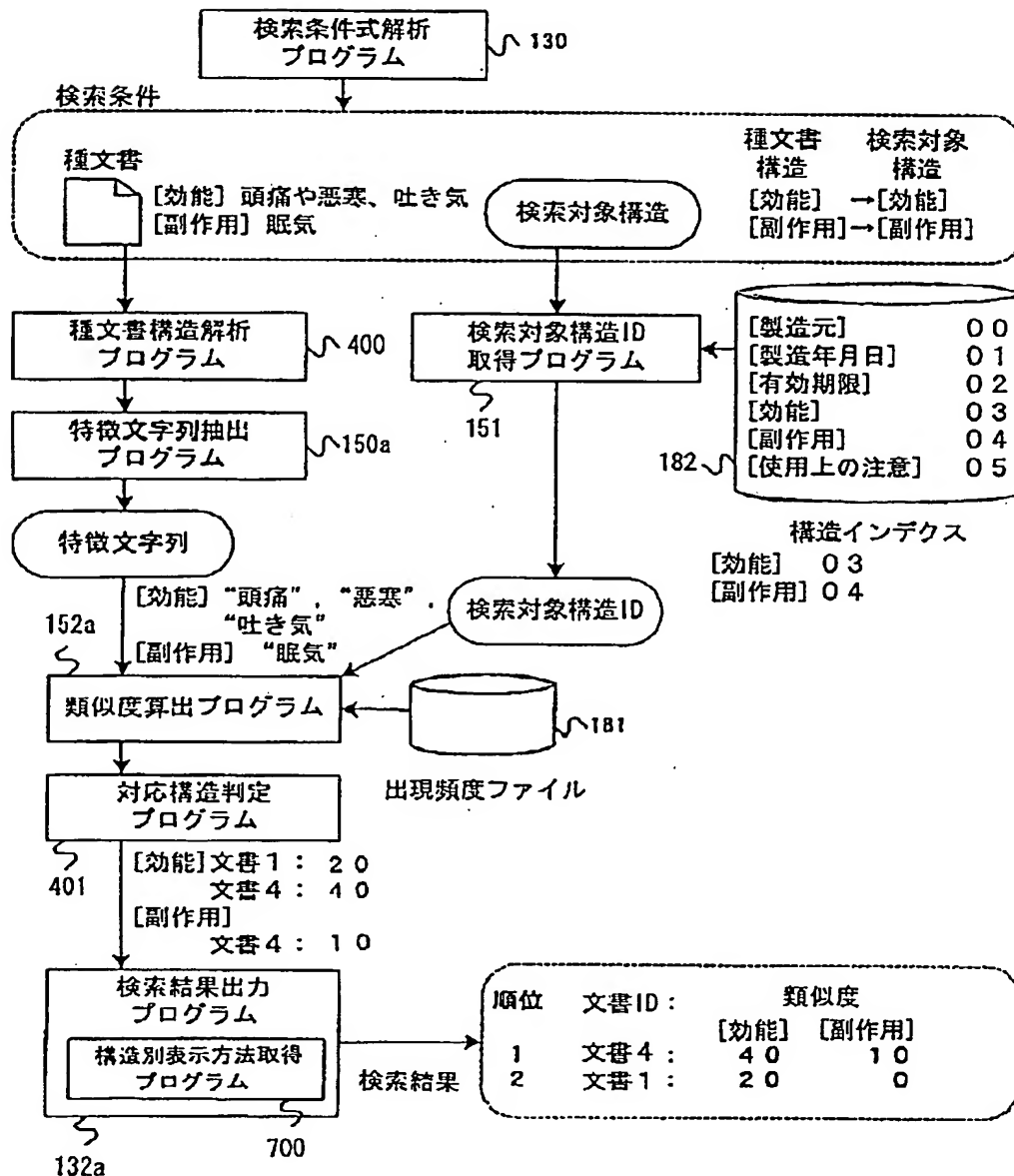
【図8】

図8



【図10】

図10



フロントページの続き

(72) 発明者 菅谷 奈津子
 神奈川県川崎市幸区鹿島田890番地 株式
 会社日立製作所システム開発本部内

(72) 発明者 稲場 靖彦
 神奈川県川崎市幸区鹿島田890番地 株式
 会社日立製作所システム開発本部内

(72) 発明者 山口 明彦
 神奈川県川崎市幸区鹿島田890番地 株式
 会社日立製作所システム開発本部内

(72) 発明者 後地 陽介
 神奈川県横浜市戸塚区戸塚町5030番地 株
 式会社日立製作所ソフトウェア事業部内

F ターム (参考) 5B009 QA09 VA02
5B075 ND03 NK06 NK39 PP13 PQ02
PQ22 PQ36 PQ46 PQ75 PR06
QM08